4-23-2021

Modeling the mechanistic behavioral logic supporting adherence to hormone therapy in breast cancer patients

Gina Kersey
gek9118@rit.edu

RIT

Modeling the mechanistic behavioral logic supporting adherence to hormone therapy in breast cancer patients

by

Gina Kersey

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Committee: Dr. Gary Skuse, Dr. Gordon Broderick, and Dr. Matt Morris

School/Department of Thomas H. Gosnell School of Life Sciences

College of Science

Rochester Institute of Technology

Rochester, NY 14623-5603

April 23, 2021

Abstract

This project proposed that given a basic model of the decisional logic driving stable adherence it might be possible to reliably anticipate increased vulnerability to discontinuation and use such a mode to design behavioral interventions that are specifically tailored to deliver course corrections that are temporally and contextually optimal. A logistic regression was coded and run using the core model predicted by the experts as a base for how the variables interact to make predictions. Then a logistic regression with stepwise selection was coded and run using the core model as a base for how the variables interact and to allow for variables to be added to make the prediction more accurate. The logistic regression with stepwise selection produced both an augmented core model and a *de novo* model. There were two nodes that had a variable added over 75% of the time to the augmented core. These nodes were health literacy and treatment fatigue. The accuracy for the core model, augmented core model, and *de novo* model were all accurate with respective overall accuracies being 63.5%, 66.3%, and 67.4%. However, it is important to note that we were able to outperform the expert model by doing the logistic regression with stepwise selection. The results show that the causal interaction diagram predicted by the experts was fairly accurate. These results then can be used to further the research needed to make a program that will help predict the chances of adherence and to be able help doctors work with the patients to stay adherent.

Introduction

Breast cancer is the most common type of cancer in women, there are about 1,200,000 new cases of breast cancer each year worldwide.[Lumachi et al., 2013] Though there are several different subtypes of breast cancer [Russnes et al., 2017] distinguished by histological and

molecular profiles. The current project focuses on estrogen receptor positive (ER+) breast cancer, a molecular subclass capturing over 70% of breast cancers.

Adjuvant endocrine therapy for stage I-III estrogen receptor positive (ER+) breast cancer provides a substantial survival benefit[EBCTCG, 2011]. Evidence suggests that despite the benefits of adherence to adjuvant hormonal treatment (ET) for the full duration, in the first year discontinuation of adjuvant ET is between 7% to 14% and that may increase to between 31% to 60% by the end of the fifth year. [Winer 2005, Moore, Hadji 2010, Ma 2008] Some of the benefits of sustained adherence to ET include an increase in the overall chance of survival and reduced cancer specific mortality, decreased risk of recurrence, and decreased risk of contralateral breast cancer, which is finding a tumor in the opposite breast of the first cancer.[Burstein 2014] There are also a number of risks associated with ET as well as significant deterrents to continued adherence. Though relatively rare, such risks vary from one therapeutic agent to another and can include developing endometrial cancer, menopausal symptoms, deep vein thrombosis, pulmonary embolism, ischemic heart disease, osteopenia/osteoporosis, and uterine cancer.[Chay et al., 2016] Other significant drivers of discontinuation of ET are the significant side effects which include hot flushes, muscle and joint pain, weight gain, fatigue, depression, difficulty concentrating, numbness or tingling in the extremities, vaginal dryness, and hair loss.[Paranjpe et al., 2019]. More recently, Shinn et al. [2019],reported on changes in the patterns of concerns over the first 5 years of ET in 216 individuals and found that leading concerns or behaviors driving discontinuation such as forgetfulness, cognitive fatigue, and worry over treatment cost evolved over time. The results that Shinn et al. found bring up important questions of how would you intervene early in treatment and late in treatment. Also, what variables would change based on at what point of treatment the patient is.

There are many different therapies that require adherence to a specific treatment regimen whether it is a medication or specific exercises. Simpson et al. [2006] studied the connection between adherence to drug therapy and mortality. They found that good adherence for patients that were adhering to either placebo or beneficial drug therapy had half the risk of mortality than those with poor adherence. However if the drug proved harmful, the patients with good adherence had a higher mortality rate than those with poor adherence. [Simpson et al., 2006]. Narayanan et al. [2017] studied adherence to cystic fibrosis therapies and the effects of not adhering to treatment. The authors found that for most patients the adherence was subpar, but this did vary according to the treatment, age of the patient, and even the season. They found that when patients are adherent to cystic fibrosis therapies, they can experience a lower disease burden and improve the patient outcomes. However, in patients who did not have good adherence it was seen that there was a clinical and economic burden of the disease because they had to receive more treatment later on as a result of not being adherent to their treatment. Another study of the adherence of oral anticoagulation therapy found that for one treatment there was an adherence level of 40% and for the other treatment there was an adherence of 47% [Chen et al., 2020]. Oates et al. [2019] studied the objective versus self-reported adherence rates in airway clearance therapy in cystic fibrosis. They found that while there was a mean adherence of 61% in the objective findings, the amount of highly adherent subjects was 31% and 28% were low adherent. The self-reported adherence levels showed 65% of subjects reported themselves as high adherence and only 8% of them reported low adherence. [Oates et al., 2019] This overestimated actual adherence could affect other treatments not just cystic fibrosis.

The study done by Chen et al. [2020] found that there were many variables that interacted with the patient's decision to discontinue use of the medication or have varying levels of

adherence. Ghembaza et al. [2014] asked if the patient's knowledge of hypertension complications on adherence to therapy impacted the patient's adherence to treatment. While this is not the same disease that my project is studying, there could be reasons for staying adherent or becoming non adherent that are not specific to one disease. Ghembaza et al. [2014] found that there is a positive relationship between what is known about hypertension complications and adherence to treatment. [Ghembaza et al., 2014]. There have also been studies into predicting if a patient will stay adherent to therapy. Essery et al. [2016] found a number of factors that influenced the patient's adherence to home-based physical therapies. Those factors include intention to engage in the therapy, self-motivation, self-efficacy, previous adherence to exercise-related behaviors, and social support.[Essery et al., 2016]. Some of these factors could probably be applied to other therapies and social support is one that is important in most therapies including this project.

To increase adherence, the researchers must first find out why the patients are not adhering to treatment. The consequences for not adhering to the treatment plan can range from very small to quite large. One disease that has a larger consequence for non-adherence was breast cancer. It was found in breast cancer patients that those who are non-adherent have a higher mortality rate. The estimated survival for those who discontinued treatment was 73.6% after 10 years versus 80.7% for those who were adherent to treatment[Hershman et al. 2010]. They found that the survival at 10 years for those who continued the study had a survival rate of 81.7% if they were adherent and 77.8% if they were non adherent. This is not a large difference, but it does prove that those who discontinued early or were non-adherence did have an increased mortality rate. [Hershman et al., 2011]. Having a predictive program that can account for the different variables that affect adherence to treatment could be helpful with increasing adherence

to the treatment, thereby improving patient outcome. Being able to predict the next stage that a variable might be when a person is on a treatment plan can be very helpful to prevent non-adherence. It could be possible that once the program can predict the next stages of the variable accurately, that it could be developed to accurately predict if a patient will stay adherent to a medication. This could help doctors prepare for if a patient might become non-adherent and work with the patient to keep them adherent to the treatment. Keeping a patient adherent to a treatment could be beneficial to the patient. In patients with cystic fibrosis, adherence to medication can decrease cost of treatment in the long run and lower disease burden [Narayanan et al., 2017]. In patients with breast cancer, adherence to treatment leads to an increase in life expectancy over those who are not adherent[Hershman et al., 2010]. What has been found in all therapies that have been discussed so far is that staying adherent to any treatment that has been discussed so far is beneficial to the patient whether it helps them live longer or have reduced medical bills or just simply have a better quality of life.

Currently, the underlying behavioral processes that govern changes in the relative importance afforded to these concerns as well as their role in driving an upcoming discontinuation of therapy are poorly understood. We propose that given a basic model of the decisional logic driving stable adherence it might be possible to reliably anticipate increased vulnerability to discontinuation and use such a model to design behavioral interventions that are specifically tailored to deliver course corrections that are temporally and contextually optimal.

Approaches for extracting such mechanisms or decisional processes from data *de novo* include statistical methods based on conditional probabilities such as those initially proposed by Pearl [2003] however such methods require substantial amounts of numerical data to infer causal relationships with high confidence. This project does not have a large amount of numerical data

which could make the methods that Pearl discussed more difficult to apply because those methods are most successful with lots of data. Another approach more suitable to limited data consists of proposing a set of causal decisional processes, formulating these into a logic network and comparing the predicted dynamic responses against observed behaviors. This project would work off of a logical framework similar to what Abou-Jaoude et al. did[2016]. They started with a logical framework and then defined models and their dynamics where most common variants are presented focusing on updating schemes and their impacts on dynamical properties. With a logical framework it is important to take the updating schemes and their impacts on dynamical properties into account to create an accurate prediction. If the model does not take into account the dynamics of the model and the data into account when making a prediction the prediction will not be as accurate as it could be [Abou-Jaoude et al., 2016]. Toole et al. discuss using a logic model applied to depression that can then be used by a computer program to help make better predictions or computer simulations[Toole et al., 2018 depression]. Using a preliminary logic model can give a program a starting point and depending on how accurate the logic model is it could help the program to make more accurate predictions. Toole et al. used an optimization-based trial and error to discover novel relationships between well-being and dopamine and acetylcholine. This allowed for the authors to be able to make a specific feedback network that would promote well-being by dopamine and acetylcholine levels and promoted norepinephrine while inhibiting cortisol [Toole et al., 2018 optimization].

Research Question: How would I intervene early in treatment and late in treatment? What variables would change based on what point of treatment the patient is at?

The overarching aim of this study was to make a model that predicts the decision to discontinue treatment based on data from a pilot study. Within that aim, we set out to complete a decisional network model by comparing the core relationships proposed by a domain expert to those extracted from the data using a simple logistic regression with stepwise selection based approach.

Methods

Data.

The data that has been collected on many different variables that may impact the continuation of therapy, such as cost, side effects, sense of urgency, and others that affect adherence to adjuvant ET. The data that was collected can be used to look for trends and determine why people were or were not adhering to the treatment.[GriffinS, 2006]

Events resulting in a change of adherence status were of specific interest. A total of 26 changes in status were observed, namely 5 decisions to resume adherence and 21 decisions to discontinue. These occur at various phases of ET as outlined in Table 1. Timepoints 9->11 and 11->13 are when most switches to nonadherence occurred. Many more patients became nonadherent and there were very few that became adherent.

Table 1: Changes in Adherence

Timepoint	Became Adherent	Became Nonadherent
1 → 3	0	2
3 → 5	0	2
5 → 7	0	1
7 → 9	2	2
9 → 11	1	7
11 → 13	0	5
13 → 15	1	2
15 → 17	1	0
Totals	5	21

Table 1: Adherence changes over time during adjuvant ET. 55/82 subjects entered adherent and stayed adherent, 4/82 subjects entered non-adherent and stayed non-adherent, 21/82 subjects had at least one change in adherence. There was a total of 26 changes in adherence.

The discretized data needed to be rearranged in a way that the regression would be able to predict the next time point for the variables from the current time point. For each subject the data had to be set up so within each subject the timepoints go in pair of two from the first time point to the next time point, for example time point 1 to 3 then timepoint 3 to 5. The first time point is listed as the original names of the variable and the second time point has _y added to the end of the variable names. It is important to note that in Figure 1, there is a node called adherence which refers to adherence from timepoint x to timepoint y not the overarching adherence to the medication.

Computational Approach

The data were interrogated through a program that performed probabilistic predictions of the next state based on data collected at timepoints preceding the current one. The basic forecasting model was a logistic auto-regressive moving average with exogenous variables or ARMAX. Both the logistic regression and logistic regression with stepwise selection were set up to predict the next timepoint based on the previous and current timepoints and the goal was to accurately predict the state of the variable in the next timepoint.

There were three basic sets of exogenous regressor variables used. First a core model was proposed by domain experts describing their understanding of cause and effect relationships driving adherence. The core model details the variable and the variables that are proposed to have negative or positive causal interactions with that variable. Second, this core model was augmented with new statistically inferred interactions using a logistic regression with stepwise selection procedure. This second model starts with the core model and then alternates between adding and removing new variables until the new variable set is statistically significant in reducing prediction error. The last model is a *de novo* logistic regression with stepwise selection model. The *de novo* model is created using logistic regression with stepwise selection, but starts with a blank slate where all of the measured variables are candidates for selection.

The first step to making the logistic regression was to make a multivariate logistic regression on one of the variables to get used to the program that was being used and to learn how to have the program allow for multiple variables to interact with the one that was currently being predicted. Once this was working a multivariate logistic regression was written to run on each node taking all variables that are predicted to affect the node into account. The training/test

split used for the logistic regression was 30 training, 70 test. The logistic regression was coded in Python and the sklearn library was used. Using the sklearn library allowed me to use the logistic regression method that is built into the sklearn library.

The logistic regression algorithm with stepwise selection [Rawlings, 1998] was applied at each node in turn and new input variables added such that the next state of that node was predicted with minimal error. The partial F null probability for inclusion of a new term was set to $p = 0.05$ for elimination of an existing term. The logistic regression with stepwise selection was set up in a way that allows for the program to be able to find the highest accuracy for each node. The logistic regression with stepwise selection was also set up to be able to be run with either the expert predicted interacts as a starting point for what variables the logistic regression with stepwise selection should include or with no base suggestions. The logistic regression with stepwise selection added variables that helped the program make the highest prediction accuracy for the nodes. The training/test split for the logistic regression with stepwise selection was a 30 training/70 test split like what was done with the logistic regression. The logistic regression with stepwise selection was programmed in Python and the library sklearn was used for this program as well. The logistic regression method from sklearn was used to run the logistic regression with stepwise selection as well because to find the highest prediction a logistic regression was run on the node and then the logistic regression with stepwise selection part allows for the logistic regression to continue being run on one variable until the highest prediction is found.

Figure 1: Causal interaction diagram

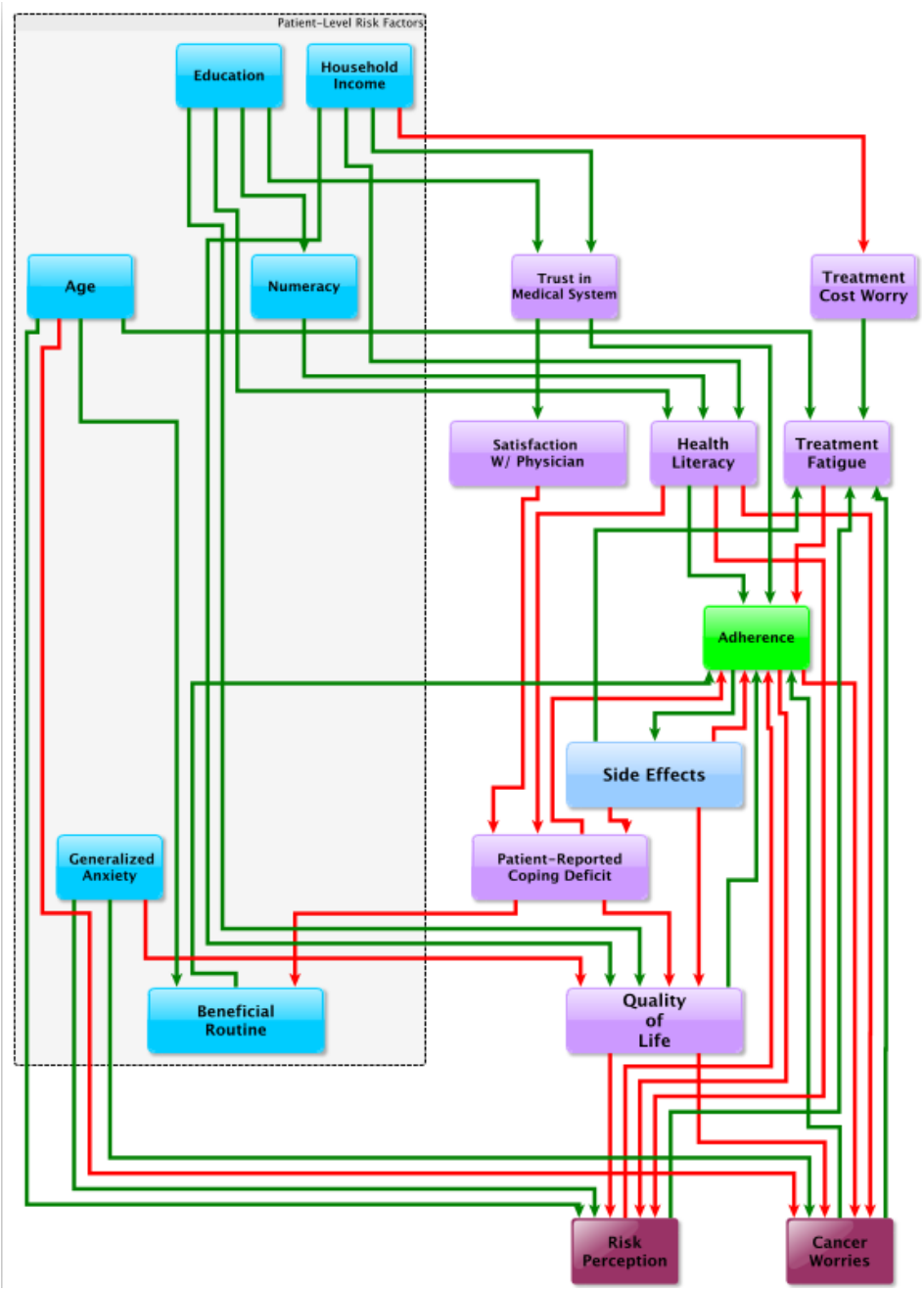


Figure 1: Causal interaction diagram. This circuit diagram describes the decisional logic flow across all the variables, how they interact with each other, how they can influence

the overall quality of life and how they influence the patients adherence to the therapy. (Shinn et al., 2020).

A basic logistic regression was run against the basic structure of the expert-informed core model to determine its accuracy in reproducing the experimental observations. This was done by programming a core set of variables for each node according to the connectivity described in Figure 1. The logistic regression used the measured node state and that of its upstream input nodes at the current timepoint to predict the node's state at the next timepoint. The data was randomly split into two sets, 70% of observations for testing and 30% of observations for training. These training and test sets were repeatedly subsampled and the logistic regression was applied 100 times.

Similarly, test and training subset selection and the logistic regression with stepwise selection were also repeated 100 times in the case of the augmented core model and the *de novo* model.

A comparative analysis of the model accuracies produced in each case was performed to determine if the augmented core model and the *de novo* models improved prediction accuracy on average over that obtained by using the expert-informed variant. Null probability p-values were computed with a threshold for significance of 0.002 to determine if there was a significantly higher average accuracy from the logistic regression with stepwise selection models compared to the expert-informed logistic regression results. Likewise, the overall frequency of selection for each of the candidate input variables was analyzed to identify new upstream modes added to the core model 75% of the time or more.

All of the calculations for the comparison analysis were performed in Python. The comparison was done using a t-test as well as the p values and for the t-test the library scipy.stats was used.

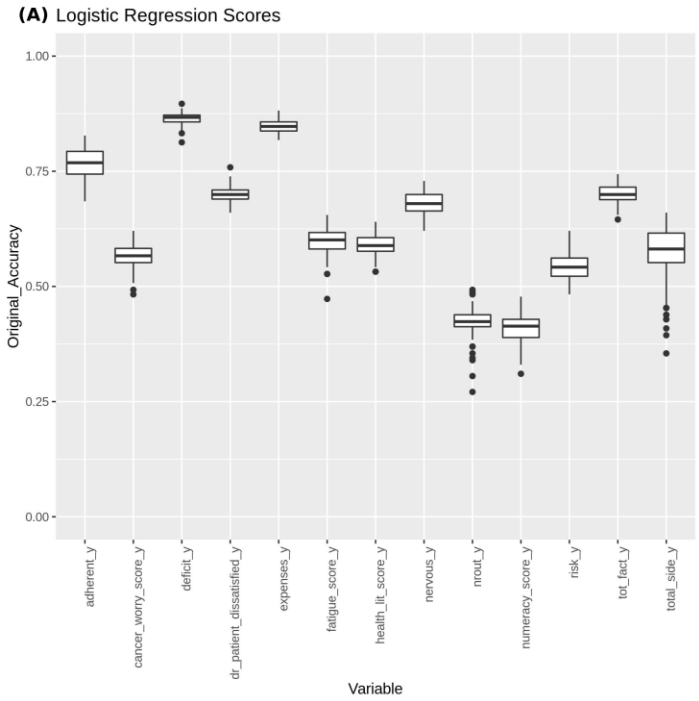
Results

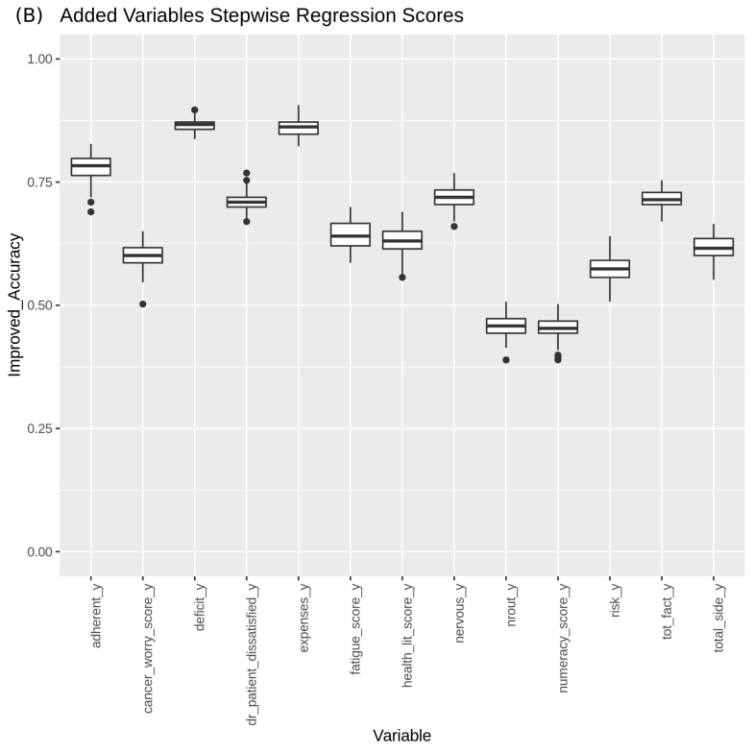
Evaluating prediction accuracy an expert-informed model and its variants

Using the basic connectivity at each node as described in Figure 1 and applying a logistic regression as a state transition function to predict the next state we find that only 3 nodes deliver an accuracy above 75%. These are the adherence status, treatment cost worry, and patient-reported coping deficit. Given the expert-informed model structure the next state of most nodes is predicted with an accuracy ranging between 50% and 75%, with the exception of numeracy, and beneficial routine for which prediction accuracy is especially poor (<50%) (Figure 2A). The addition of novel edges to the core expert model as inferred from the data using a stepwise selection routine improved the accuracy of most nodes raising the overall average accuracy from 63.53% to 66.34% (p=0.05**)**. The most noticeable increase in average accuracy was observed for nodes treatment fatigue, adherence, risk perception, cancer worry, health literacy, numeracy, quality of life, side effects, general anxiety, patient satisfaction with doctor, treatment cost worry, and beneficial routine. All of these nodes had a p value greater than 0.05. (Figure 2B). Finally, selecting upstream mediators naively from the data alone, in the absence of any prior structure, had the effect of normalizing the accuracy across all nodes to an average of roughly 70%. Average accuracy in the structurally naive model increased for some nodes and decreased for others but for all nodes the range of accuracy values across the 100 repeated subset selections increased noticeable from sigma ~ 63.53% to sigma ~ 67.39%. In figure 2C you can see that the boxes on the graph are much wider than graphs 2A and 2B. This suggests that the accuracy is

dependent on the subset selection and that without a base model to go off of the model is not that stable.

Figure 2: Predictive accuracy of network variants





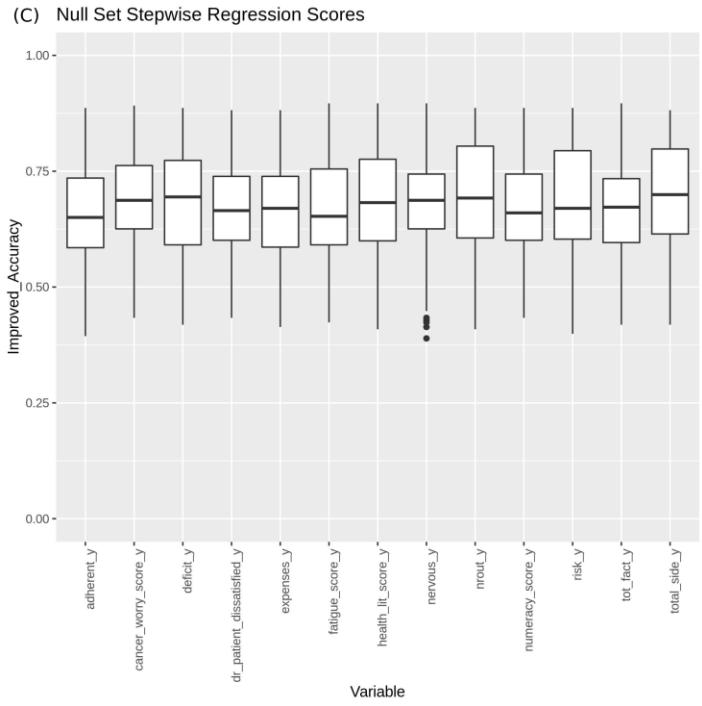
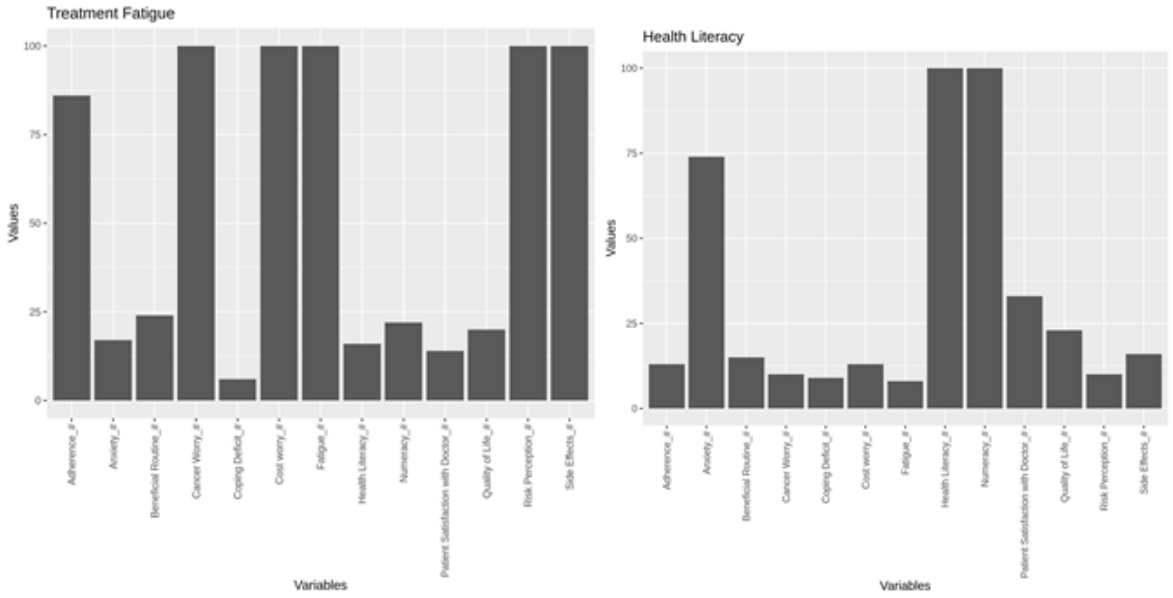


Figure 2: Accuracy in predicting the next state for individual nodes using a logistic regression applied to an expert-informed model structure (A), a stepwise selection of upstream

interactions to be added to the core expert model (B), and a *de novo* stepwise selection of upstream from all model nodes (C).

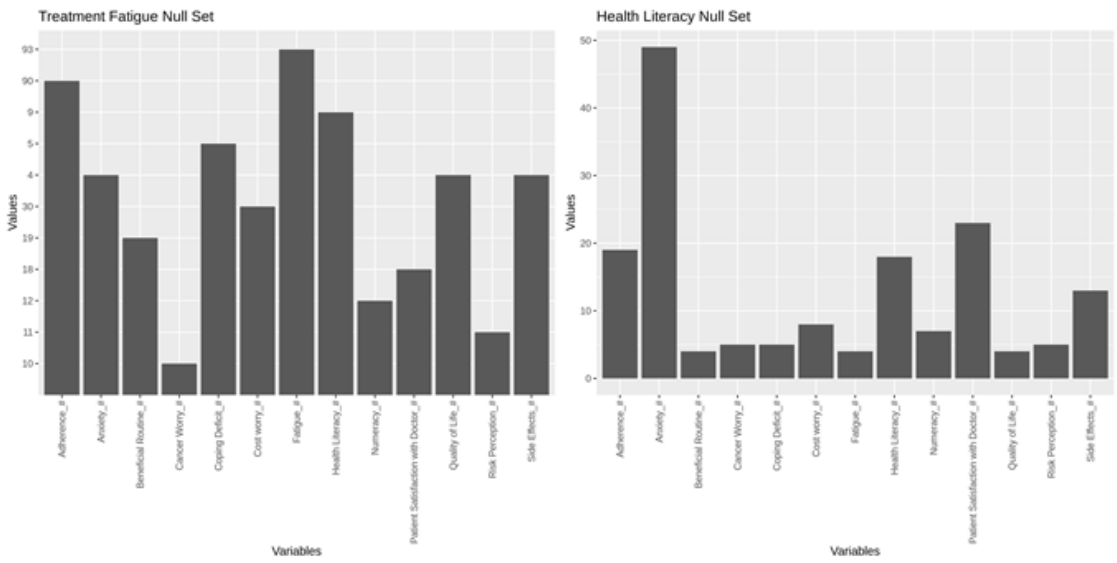
Figure 3: Consensus voting of augmented core model structure



(A) **(B)**

Figure 3: Two graphs that show a significant node that had a variable added to it during the augmented core model run a significant amount of times. The A graph shows the significant node is treatment fatigue and the variable adherence was added by the regression about 90 times out of the 100 runs. The B graph shows the significant node is health literacy and the variable that was added a significant amount of time is anxiety which was about 75% of the time when it was run.

Figure 4: Consensus voting of *de novo* model



(A) **(B)**

Figure 4: Two of the significant nodes from the *de novo* model. Graph A represents the treatment fatigue node and it showed that the adherence variable was added over 90% of the time. The graph B shows the node for health literacy and the variable anxiety was added to the prediction about 50% of the time.

Figure 3 focuses on two nodes that had a variable added the majority of times to the augmented core model out of the 100 runs that were done. Figure 4 focuses on the same two nodes from Figure 3, but it is looking to see if the variables that were added the majority of the runs to the augmented core model were also added during to the *de novo* model. The two nodes that are looked at in these figures are health literacy and treatment fatigue. In figure 3A you can see that general anxiety is added the majority of the runs to the node health literacy and in figure 4A general anxiety is added to the health literacy node about 50% of the time which is more than any other variable for that node. In figure 3B you can see that the variable, adherence, was added

the majority of the time to the node, treatment fatigue and in figure 4B adherence was also added the majority of the runs to the treatment fatigue node.

Table 2: The average accuracy score for each node.

Nodes	**Core Model**	**Augmented Core Model**	***De novo* Model**
Treatment fatigue	0.5985	0.6448	0.6653
Adherence	0.7671	0.7786	0.8126
Risk Perception	0.5409	0.5751	0.5982
Cancer Worry	0.5666	0.6016	0.6253
Health literacy	0.5890	0.6302	0.6316
Numeracy	0.4095	0.4526	0.4578
Quality of Life	0.7015	0.7160	0.7349
Side Effects	0.5734	0.6140	0.6137
General Anxiety	0.6800	0.7184	0.7187
Patient Satisfaction with Doctor	0.7010	0.7098	0.7130
Cost Worry	0.8469	0.8603	0.8582
Beneficial Routine	0.4213	0.4575	0.4655
Coping Deficit	0.8635	0.8653	0.8659

Table 2: Average accuracy score of each node. The first column labeled “Core Model” which is the average accuracy of each node when the data was run through the logistic regression. The second column “Augmented Core Model” which is the average accuracy of each node when run through the logistic regression with stepwise selection. The third column “*De novo* Model” is the average accuracy of each node when run through the logistic regression with stepwise selection.

Of the 13 nodes re-evaluated by applying stepwise feature selection to augment the expert core model, 12 showed significantly improved average prediction accuracy with a less than 0.05 p value. These 12 nodes were treatment fatigue, adherence, risk perception, cancer worries, health literacy, numeracy, quality of life, side effects, general anxiety, patient satisfaction with doctor, treatment cost worry, and beneficial routine.

In the *de novo* model assembled without prior knowledge of structure we also found 11 out of 13 nodes that improved significantly in accuracy compared to the logistic regression model. The 11 nodes that are significant are treatment fatigue, adherence, risk perception, cancer worries, health literacy, numeracy, quality of life, side effects, treatment cost worry, beneficial routine, and coping deficit. In the *de novo* model there were also 3 nodes that had a 5% or more increase in accuracy. Those nodes were treatment fatigue, risk perception, and cancer worry.

Figure 5: Updated causal interaction diagram

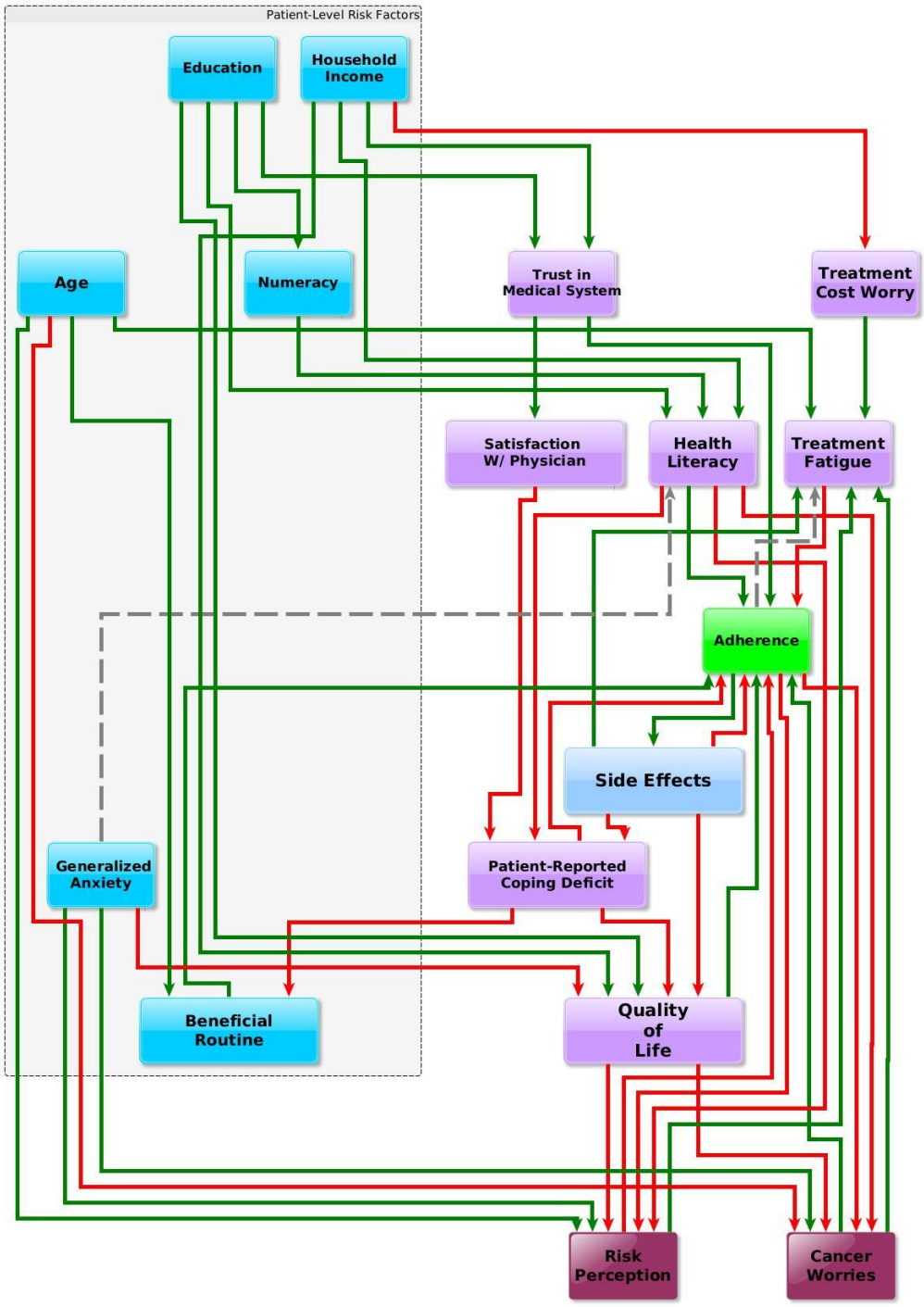


Figure 5: The two new novel edges were added with gray arrows to the original causal interaction diagram to make the updated causal interaction diagram. One of the new novel edges

is generalized anxiety's connection to Health literacy and the other is adherence's connection to treatment fatigue. These are the variables that were added the majority of the times during the logistic regression with stepwise selection.

The two connections that were added in figure 5 were determined based on what was seen in figure 3 and figure 4. The accuracy for the node treatment fatigue increased quite a bit during the logistic regression with stepwise selection for both the augmented core model and the *de novo* model and the logistic regression with stepwise selection added the variable adherence when it was predicting the outcome for treatment fatigue. When generalized anxiety was added to the node health literacy by the logistic regression with stepwise selection to both the augmented core model and the *de novo* model, there was an increase in accuracy seen.

Model stability and contribution of prior knowledge

The variability in the *de novo* model is quite a bit larger than that seen in the core model and the augmented core model, seen in Figure 2. Variability increased for a few different nodes. The variability increased for the nodes treatment fatigue, risk perception, health literacy, numeracy, quality of life, side effects, patient satisfaction with doctor, and cost worry. The standard deviation for treatment fatigue for the *de novo* model was 0.0343 while the standard deviation for the augmented core model was 0.0254. For the node, risk perception, the standard deviation of the *de novo* model was 0.266 and the standard deviation of the augmented core model was 0.255. The standard deviation of health literacy of the *de novo* model was 0.0288 and the standard deviation for the augmented model was 0.266. The node numeracy had the standard deviation of the *de novo* model was 0.248 and the augmented core model was 0.214. 0.0190 was the standard deviation of the node quality of life for the *de novo* model and 0.0186 was the

standard deviation for this node for the augmented core model. The node, side effects, when looking at the standard deviation from the *de novo* model it was 0.2655 which is higher than the standard deviation of the augmented core model was 0.259. The node, patient satisfaction with doctor, had a standard deviation of 0.0189 in the *de novo* model and this is an increase from the standard deviation of 0.0173 in the augmented core model. Coping deficit was the last node that had an increase in variability from the augmented core model to the *de novo* model when looking at the standard deviation. The standard deviation for the *de novo* model was 0.139 and the standard deviation for the augmented core model was 0.123. The standard deviations are fairly similar, but there is more variability seen in many of the nodes in the *de novo* model than the variability seen in the augmented core model.

Discussion

The experts that created the putative network model were correct about most nodes and which variables interact with each node. There were a couple nodes that had variables added consistently when the logistic regression with stepwise selection was run. One node where that happened is the treatment fatigue node which had adherence added to it almost every run. The other node where consistent improvements were made is the health literacy node which had general anxiety added to it almost every run. Overall the expert opinions were quite accurate and were a very good starting point when it came to predicting the next timepoint for that node.

When you look at the average of all nodes when running the logistic regression with stepwise selection for either the *de novo* or the augmented core model the average is higher than when the logistic regression is run with the core model. The *de novo* model had an average node score of 67.4% and the augmented core model had an average node score of 66.3% which is

compared to the average node score of 63.5% from the logistic regression. There were some nodes that had a very high accuracy for each method and then there were some that had a consistently low accuracy for each method. When considering the overall goal of this project it is important to have as high a possible accuracy for each node. This is because it will help with a better performance in the final program that could be used in a clinical setting.

Results presented in Figure 2, that both the core model and the augmented core model produced a relatively consistent accuracy from one subsample of data in predicting the next state for each individual node. However, we observed a broad disparity in the average prediction accuracy from one node to the next. The opposite was true of the *de novo* assembled model where a much broader range of accuracy values at each node was obtained from one data subsample to the next while the average performance across nodes was much more uniform.

The augmented core model and the core model nodes in Figure 2 do not have much variation between the accuracy of each node. For the core model that was expected because the accuracy would not be expected to change much since the variables stay the same the whole time. The augmented core model also did not change much because for most of the nodes the predicted core model from the experts were accurate.

One frequent reason for that can be traced back to changes in the coordinated patterns exhibited in each of the subsampled data sets. For this project, we are trying to find a universal model for all subjects, but this type of subset specific results could suggest that it might be more appropriate to stratify the subjects and have multiple models which could be either one per subject or data subset. In addition, to make this even more complicated the regressor variables are not statistically independent - they are more or less coupled as they are all members of a

biological regulatory network and as such their measurements will be expressed in specific patterns. These patterns may shift significantly from one data subset to another affecting the choice of model terms at each step. However, when we look at the core model and the augmented core the accuracy values appear quite consistent and vary over a very narrow range. This could mean that the data subsets might also be reasonably consistent. This points to the model structure itself and the degrees of freedom with which new terms can be chosen. Figure 3 and 4 show how in both the augmented core model and the *de novo* model have a variable added consistently to a node. The nodes that have a variable added consistently during the runs in both the augmented core model and the *de novo* model are the health literacy node and the treatment fatigue node. There were only two nodes that had a variable added consistently which suggested a broad range of statistically equivalent solutions [Rawling, 1988]. By setting a priori a number of initial input variables, the choices available for new terms were limited, which decreases the degrees of freedom, and produces a more narrow family of models since most of the model is predetermined. This seems to be a more likely culprit.

The health literacy node had the variable general anxiety added the majority of the runs for both the augmented core model and the *de novo* model. The treatment fatigue node had the variable adherence added to the majority of the runs for the augmented core model as well as the *de novo* model. Having these nodes be added more than 75% of the time in the augmented core model and about 50% of the time in the *de novo* model means that these variables make it more likely to cause the prediction to be more accurate. Health literacy had general anxiety added the majority of the runs and these suggest that node general anxiety interacted with the node health literacy and when included in the model for that node it led to better predictions for the next timepoint. The same thing is true for the node treatment fatigue and the augmented core model

added the node adherence the majority of the time. This most likely means that the node adherence has some interaction with the node treatment fatigue and when it is included in the model it increases the number of correct predictions for the next timepoint of treatment fatigue.

There are a few limitations that could occur with this project. One of the limitations is that the sample size of the data is small which means that it could be difficult to use probability to reliably predict any changes in drug adherence. Though incomplete these models can serve as a basis for adjusting the recruitment and assessment protocols to focus on higher risk individuals such that the information content of new data is increased in a specific and premeditated way [Vashishtha et al., 2015; Vildela et al., 2015] - picking specific variables to focus and to use studies on similar data to make specific assumptions about what data variables are most important to focus on. There will have to be a specific type of distribution assigned to the data and the specific assumptions will be used to determine how to make the Petri nets based off of studies that have been done on similar types of data. [Rawlings J, 1998]

While there were a few limitations to take into account during this project, the program was still accurate for many of the nodes. Using a logistic regression with stepwise selection allowed for the discovery for a couple nodes that benefited from having another variable added. Other than the couple nodes that benefited from a variable being added, the expert made causal interaction diagram that was quite accurate and allowed the logistic regression to make accurate predictions for most variables. One variable that had fairly low accuracy was beneficial routine. There could be multiple reasons for that, one of which could be that there is not enough knowledge about what can affect that node and the variable that would affect it is not currently one of the variables that is part of the model. Another factor to consider is that to increase the accuracy of some of the nodes there would have to be more timepoints for the node. The current

study used a questionnaire every six months, but some of these variables might have had a higher accuracy if the questionnaire was every week, but that is not something that can be done in a clinical setting.

Conclusion

Though the addition of new data-informed relationships did improve the predictive performance of the state transition function at least somewhat at a majority of nodes the basic expert-informed model nonetheless performed surprisingly well. Moreover, the inclusion of this baseline prior knowledge as a core model scaffold served to reign in the high degrees of freedom and reduce the range of solutions compared to the completely naive *de novo* approach. This is especially important in cases where there is little data to provide such constraints and where highly interrelated variables amplify the number of statistically equivalent models. Further improvements are needed before accurate predictions of adherence to therapy can be made. Additional data would further strengthen the validation and the current models could support the design of information recruitment strategies.

Predicting the next timepoint of each individual node is the first step in creating a program that is able to accurately predict the chance of adherence to the treatment. Having a program that is able to accurately predict the chance of adherence from one timepoint to the next could be useful in a clinical setting. The program could be used in a clinical setting by entering in the patient's scores for each variable and then using the program to predict if they will continue to be adherent to the treatment. Then using the results from the program the patient's doctor would be able to discuss with them reasons why they might not continue treatment and the doctor could try to work with the patient so they continue the ET.

While this project focused specifically on making a model that was able to make predictions in relation to variables thought to affect ET, this thinking and application of machine learning could be applied to other treatments for cancer and other diseases as well.

References

1. Lumachi, F., Brunello, A., Maruzzo, M., Basso, U., & Basso, S. (2013). Treatment of Estrogen Receptor-Positive Breast Cancer. *Current Medicinal Chemistry, 20*(5), 596-604. doi:10.2174/092986713804999303
2. Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A., & Caldas, C. (2017). Breast cancer molecular stratification. *The American Journal of Pathology, 187*(10), 2152-2162. doi:10.1016/j.ajpath.2017.04.022
3. EBCTCG: Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The Lancet.* 2011.
4. Winer E, Hudis C, Burstein H, et al.: American Society of Clinical Oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for postmenopausal women with hormone receptor-positive breast cancer. *J Clin Oncol* 2005 *23:*619-629.
5. Moore S: Nonadherence in patients with breast cancer receiving oral therapies. *Clinical Journal of Oncology Nursing.* 2010, *14:*41-47.
6. Hadji P: Improving compliance and persistence to adjuvant tamoxifen and aromatase inhibitor therapy. *Clinical Reviews in Oncology/Hematology* 2010, *73:*156-166.
7. Ma A, Barone J, Wallis A, et al.: Noncompliance with adjuvant radiation, chemotherapy or hormonal therapy in breast cancer patients. *American Journal of Surgery.* 2008, *196:*500-504.
8. Burstein, H. J., Temin, S., Anderson, H., Buchholz, T. A., Davidson, N. E., Gelmon, K. E., . . . Griggs, J. J. (2014). Adjuvant Endocrine Therapy for Women With Hormone Receptor–Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. Journal of Clinical Oncology, 32(21), 2255-2269. doi:10.1200/jco.2013.54.2258
9. Chay, Z.E., Goh, B.F., & Ling, M.H. (2016). PNet: A Python Library for Petri Net Modeling and Simulation. *Advances in Computer Science : an International Journal, 5*, 24-30.
10. Shinn EH, Broderick G, Fellman B, et al. Simulating Time-Dependent Patterns of Nonadherence by Patients With Breast Cancer to Adjuvant Oral Endocrine Therapy. *JCO Clin Cancer Inform*. 2019;3:1-9. doi:10.1200/CCI.18.00091
11. Paranjpe, R., John, G., Trivedi, M., & Abughosh, S. (2018). Identifying adherence barriers to oral endocrine therapy among breast cancer survivors. *Breast Cancer Research and Treatment, 174*(2), 297-305. doi:10.1007/s10549-018-05073-z
12. Simpson, S. H., Eurich, D. T., Majumdar, S. R., Padwal, R. S., Tsuyuki, R. T., Varney, J., & Johnson, J. A. (2006). A meta-analysis of the association between adherence to drug therapy and mortality. *BMJ, 333*(7557), 15. doi:10.1136/bmj.38875.675486.55

13. Narayanan, S., Mainz, J. G., Gala, S., Tabori, H., & Grossoehme, D. (2017). Adherence to therapies in cystic fibrosis: A targeted literature review. *Expert Review of Respiratory Medicine, 11*(2), 129-145. doi:10.1080/17476348.2017.1280399
14. Oates, G. R., Stepanikova, I., Rowe, S. M., Gamble, S., Gutierrez, H. H., & Harris, W. T. (2018). Objective versus self-reported adherence to airway clearance therapy in cystic fibrosis. *Respiratory Care, 64*(2), 176-181. doi:10.4187/respcare.06436
15. Chen, N., Brooks, M. M., & Hernandez, I. (2020). Latent classes of adherence to oral anticoagulation therapy among patients with a new diagnosis of atrial fibrillation. *JAMA Network Open, 3*(2). doi:10.1001/jamanetworkopen.2019.21357
16. Ghembaza, M., Senoussaoui, Y., Tani, M., & Meguenni, K. (2014). Impact of patient knowledge of hypertension complications on adherence to antihypertensive therapy. *Current Hypertension Reviews, 10*(1), 41-48. doi:10.2174/1573402110011411111606 53
17. Essery, R., Geraghty, A. W., Kirby, S., & Yardley, L. (2016). Predictors of adherence to home-based physical therapies: A systematic review. *Disability and Rehabilitation, 39*(6), 519-534. doi:10.3109/09638288.2016.1153160
18. Hershman, D. L., Shao, T., Kushi, L. H., Buono, D., Tsai, W. Y., Fehrenbacher, L., . . . Neugut, A. I. (2010). Early discontinuation and non-adherence to adjuvant hormonal therapy are associated with increased mortality in women with breast cancer. *Breast Cancer Research and Treatment, 126*(2), 529-537. doi:10.1007/s10549-010-1132-4
19. Pearl, J. (2003). Statistics and causal inference: A review. *Test, 12*(2), 281-345. doi:10.1007/bf02595718
20. Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., & Chaouiya, C. (2016). Logical modeling and Dynamical analysis of cellular networks. *Frontiers in Genetics, 7*. doi:10.3389/fgene.2016.00094
21. Tory Toole, J., Rice, M. A., Craddock, T. J., Nierenberg, B., Klimas, N. G., Fletcher, M. A., . . . Broderick, G. (2018). Breaking away: The role of homeostatic drive in perpetuating depression. *Methods in Molecular Biology,* 121-144. doi:10.1007/978-1-4939-7828-1_8
22. Tory Toole, J., Rice, M. A., Cargill, J., Craddock, T. J., Nierenberg, B., Klimas, N. G., . . . Broderick, G. (2018). Increasing resilience to traumatic stress: Understanding the protective role of well-being. *Methods in Molecular Biology*, 87-100. doi:10.1007/978-1-4939-7828-1_6
23. Griffin S, Clasxton K, Hawkins N, Sculpher M: Probabilistic analysis and computationally expensive models: Necessary and required? *Value Health.* 2006, *9:*244-252.
24. Rawlings J: Model Development: Variable Selection. In: Rawlings J.O., Pantula S.G., Dickey D.A. (eds) Applied Regression Analysis. Springer Texts in Statistics. Springer, New York, NY, 1998, pp 205-234..
25. Vashishtha S, Broderick G, Craddock TJ, Fletcher MA, Klimas NG. Inferring Broad Regulatory Biology from Time Course Data: Have We Reached an Upper Bound under Constraints Typical of In Vivo Studies?. *PLoS One*. 2015;10(5):e0127364. Published 2015 May 18. doi:10.1371/journal.pone.0127364
26. Videla S, Konokotina I, Alexopoulos LG, Saez-Rodriguez J, Schaub T, Siegel A, Guziolowski C. Designing Experiments to Discriminate Families of Logic Models. Front Bioeng Biotechnol. 2015 Sep 4;3:131. Doi: 10.3389/fbioe.2015.00131. PMID: 26389116; PMCID: PMC4560026

27. Zylberberg J, Deweese M: How should prey animals respond to uncertain threats? *Front Comput Neurosci.* 2011, *25:*5-20.
28. Mandelblatt JS, Schechter C, Lawrence W, Yi B, Cullen J: The SPECTRUM population model of the impact of screening and treatment on U.S. breast cancer trends from 1975 to 2000: Principles and practice of the model methods. *Journal of the National Cancer Institute Monographs.* 2006, *36:*47-55.
29. Rawlings J: Collinearity Diagnostics. In O. Barndorff-Nielsen, P. Bicketl, W. Cleveland and R. Dudley (eds), *Applied Regression Analysis.* Pacific Grove: Wadsworth and Brooks, 1988, 273-281.
30. Wold S, Trygg J, Berglund A, Antti H: Some recent developments in PLS modeling. *Chemom Intell Lab Syst.* 2001, *58:*131-149.
31. Eriksson L, Antti H, Gottfries J, et al.: Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal Bioanal Chem.* 2004, *380:*419-429.
32. Broderick G, Craddock R, Whistler T, et al.: Identifying illness parameters in fatiguing syndromes using classical projection methods. *Pharmacogenomics.* 2006, *7:*407-419.
33. Ridgway D, Broderick G, Lopez-Campistrous A, et al.: Coarse-grained molecular simulation of diffusion and reaction kinetics in a crowded virtual cytoplasm. *Biophys J.* 2008, *94:*3748-3759.
34. Broderick G, Ru'aini M, Chan E, Ellison M: A life-like virtual cell membrane using discrete automata. *In Silico Biol.* 2005, *5:*163-178.
35. Schlender A, Alperin P, Grossman H, Sutherland E: Modeling the impact of increased adherence to asthma therapy. *PLoS One.* 2012, *7:*e51139.
36. Ben-Zvi A, Vernon S, Broderick G: Model-based therapeutic correction of hypothalamic-pituitary-adrenal axis dysfunction. *PLoS Comput Biol.* 2009, *5:*e1000273.

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Gina Kersey, a candidate for the Master of Science degree in Bioinformatics, has submitted her thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name **Date**

Gary R. Skuse, Ph.D.
Thesis Advisor

Gordon Broderick, Ph.D.

Matt Morris, Ph.D.

Feng Cui, Ph.D. 475-4115 (voice) Director of Bioinformatics MS Program fxcsbi@rit.edu